基于二代测序数据的尖吻蝮基因组的组装与注释

王心雨 [abc1]，刘力榕 [b1]，朱文标 [b]，汪诗晴 [ad]，施敏辉 [ad]，杨淑慧 [c]，卢浩荣 [b*]，曹军 [b*]
a 农业基因组学国家重点实验室，深圳华大生命科学研究院，深圳，518083，中国
b 深圳国家基因库，深圳，广东，518083，中国
c 野生动物与自然保护地学院，东北林业大学，哈尔滨 150040，中国
d 生命科学学院，中国科学院大学，北京 100049，中国

摘要

对目前已知 3,000 余种蛇类的研究可为它们的基因组进化提供有价值的见解。尖吻蝮，也被称为尖鼻蝮、百步蛇或五步蛇，是一种具有重要经济、医学和科学价值的毒蛇。其广泛分布于中国东南部和东南亚，主要用于蛇毒研究。本文采用二代测序技术，组装和注释了一个高度连续的尖吻蝮基因组。基因组大小为 1.46 Gb; 其 scaffold N50 长度为 6.21 Mb，重复序列含量为 42.81%，共注释出 24,402 个功能基因。本研究有助于在遗传水平上进一步认识和利用尖吻蝮及其毒液。

关键词: 遗传学与基因组学；动物遗传学; 进化生物学

Abstract

The study of the currently known >3,000 species of snakes can provide valuable insights into the evolution of their genomes. *Deinagkistrodon acutus*, also known as Sharp-nosed Pit Viper, one hundred-pacer viper or five-pacer viper, is a venomous snake with significant economic, medicinal and scientific importance. Widely distributed in southeastern China and South-East Asia, *D. acutus* has been primarily studied for its venom. Here, we employed next-generation sequencing to assemble and annotate a highly continuous genome of *D. acutus*. The genome size is 1.46 Gb; its scaffold N50 length is 6.21 Mb, the repeat content is 42.81%, and 24,402 functional genes were annotated. This study helps to further understand and utilize *D. acutus* and its venom at the genetic level.

Keywords: Genetics and Genomics; Animal Genetics; Evolutionary Biology

尖吻蝮（*Deinagkistrodon acutus*）是属于蛇亚目、蝰科的一种有毒蛇，常被称为百步蛇、五步蛇、长鼻蝮等（如图 1 所示）[1, 2]。其毒液主要具有血液毒性，可导致凝血功能异常并促进组织损伤、水肿、急性肾衰竭等反应发生，主要作用于肺部 [3]。尖吻蝮在中国东南部、老挝和越南北部广泛分布，因其较大的身体以及毒液而具有重要的商用及药用价值 [4, 5]。目前尖吻蝮的研究主要集中在其毒液的毒性成分、被咬伤患者的症状分析等方面，以及对蛇毒的利用进行了研究，如体外抑菌、毒液中特定蛋白具有抗血栓、抗凝血活性等 [6-

9]。高质量的基因组有助于蛇毒相关基因的发现，进而可以帮助研究人员更好地了解及利用蛇毒。

本研究基于二代测序数据对尖吻蝮基因组进行组装和注释，这些数据为蛇毒相关基因的发现及利用、更好地了解蛇的系统发育和进化提供了重要的数据支持。



图 1 杨典成拍摄的一条尖吻蝮

Figure 1. An individual of *D. acutus* photographed by Diancheng Yang.

材料与方法

样本采集与测序

从安徽省黄山市（中国）获得一条重 781 g 的尖吻蝮（NCBI: txid36307）用于基因组组装及注释。取其肝、胃、肾以及肌肉组织用于 RNA 提取，另取两份肌肉组织分别用于全基因组测序（Whole Genome Sequencing，WGS）和单管长片段序列（single-tube long fragment read，stLFR）测序前的 DNA 提取。按照刘博洋等的方案提取尖吻蝮 DNA、构建文库并进行双端测序（如图 2 所示）[10]。样本采集及相关实验流程经华大基因机构审查委员会批准（BGI-IRB E22017）。

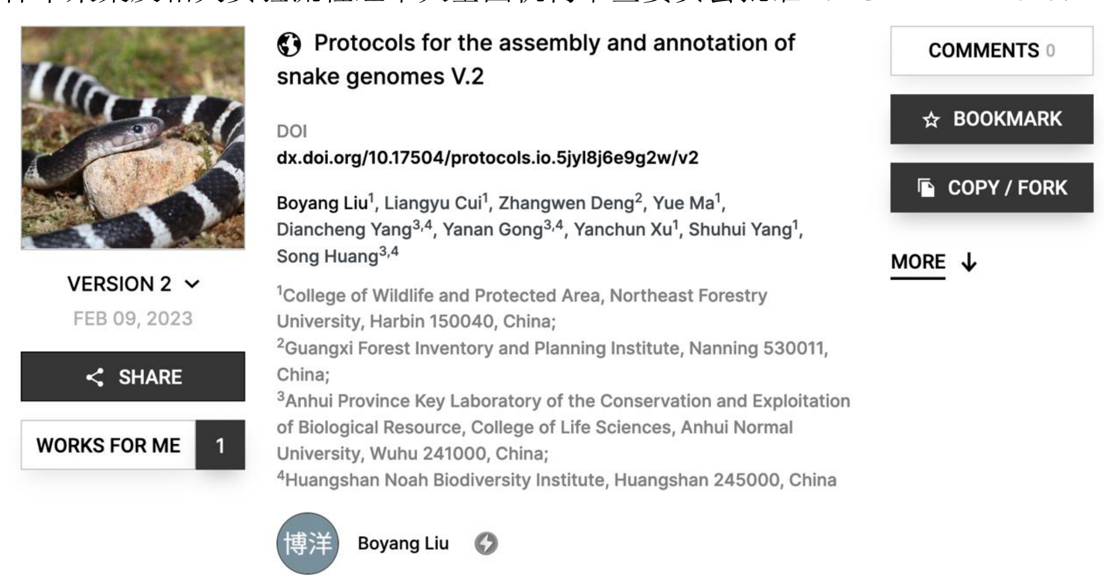


图 2 收集在 protocols.io 中用于蛇基因组一般测序的说明[10]

Figure 2. Protocol collected from protocols.io for sequencing snake genomes [10].

基因组组装 注释及评估

通过 25× WGS 测序数据评估组装的尖吻蝮基因组大小。利用 GCE（v1.0.2,

RRID:SCR_017332）的 Kmerfreq 进行 k-mer 频数统计。其输出结果表明共获取 32,372,553,516 个 k-mer 片段（k=19），将上述结果输入 GCE 并使用杂合模式（k-mer 深度峰值为 21）评估基因组大小、杂合度等 [11]。我们使用 stLFR 数据、利用 Supernova（v2.1.1，RRID:SCR_016756）进行基因组组装。为使组装的序列更完整，利用 GapCloser（v1.12-r6，RRID:SCR_015026）和 WGS 测序数据填补空白，并利用 redundans（v0.14a）去除基因组的冗余序列 [12]。使用图 2 中描述的方法获得最终基因组。我们使用从头预测和基于同源性的方法来识别基因组组装中的重复区域。基于同源性的预测使用 Blastall（v2.2.26）进行 [12]。具体而言，我们将来自 UniProt 数据库（发布号:2020_05）的东部拟眼镜蛇（*Pseudonaja textilis*）、虎斑响尾蛇（*Crotalus tigris*）、束带蛇（*Thamnophis elegans*）和虎蛇（*Notechis scutatus*）的蛋白序列比对到尖吻蝮基因组序列。按照刘博洋等描述的方案进行基因组注释和评估 [10]。

为重建系统发育树，利用 OrthoFinder（v2.3.7，RRID: SCR_017118）[13] 在中国林蛙（*Rana temporaria*，GCA_905171775.1）、古兹沙漠陆龟（*Gopherus evgoodei* ，GCA_007399415.1）、普通壁蜥（*Podarcis muralis* ，GCA_004329235.1）束带蛇（*Thamnophis elegans* ，GCA_009769535.1）、东部拟眼镜蛇（*Pseudonaja textilis* ，GCA_900518735.1）蛋白序列中寻找单拷贝同源基因。

数据验证和质控

利用 stLFR 测序产生 164.75 Gb 的主要结果文件组装了 1.46 Gb 的尖吻蝮基因组。基因组的最长和 N50 scaffold 分别为 39.38 Mb 和 6.21 Mb（如表1 所示），这表明基因组具有高度的连续性。将最终基因组与脊椎动物数据库（vertebrate_odb10）中 3,354 BUSCOs 进行比较，我们发现在尖吻蝮基因组中，3,354 个脊椎动物基因有 87.2%，即 2,924 个基因被覆盖；分别仅有 245 个（7.3%）和 185 个（5.5%）基因部分比对上及未得到比对结果。

表 1 与本研究中组装的尖吻蝮基因组相关的基因组组装数据

Table 1 Genome assembly data relative to the *D. acutus* genome assembled in this study.

Item	Category	Size
Sequencing data	stLFR (Gb)	164.75
	WGS (Gb)	96.76
	RNA-seq (Gb)	10.42
	Assembled genome (Gb)	1.46
	Longest Contig (Mb)	0.52
	Contig N50 (Mb)	0.03
	Longest scaffold (Mb)	39.38
	Scaffold N50 (Mb)	6.21
	GC content (%)	37.9

尖吻蝮基因组中重复序列总长度为 642 Mb，占整个基因组的 42.81 %（如表 2、图 3 所示）。基于从头预测，统计基因组中各种重复序列含量。最占优势的重复元件是长散在重复元件（long interspersed nuclear elements ，LINEs）（443 Mb），其次是长末端重复（long terminal repeats ，LTRs）（180 Mb）、DNAs（26.43 Mb）

和短散在重复元件（short interspersed nuclear elements ，SINEs）（0.94 Mb）。LINEs 和 LTRs 的含量分别为 29.53 % 和 11.99 %（如表 3 所示）。重复序列对于遗传信息的自我复制具有重要意义，其与物种的遗传和变异息息相关。

表 2 尖吻蝮基因组重复序列统计

Table 2 Statistics for repetitive sequences in the *D. acutus* genome.

Type	Repeat Size	% of genome
Trf	49,665,678	3.158437
RepeatMasker (RRID:SCR_012954)	254,179,490	16.16428
Proteinmask	190,282,517	12.100819
De novo	636,067,480	40.45005
Total	673,253,494	42.814856

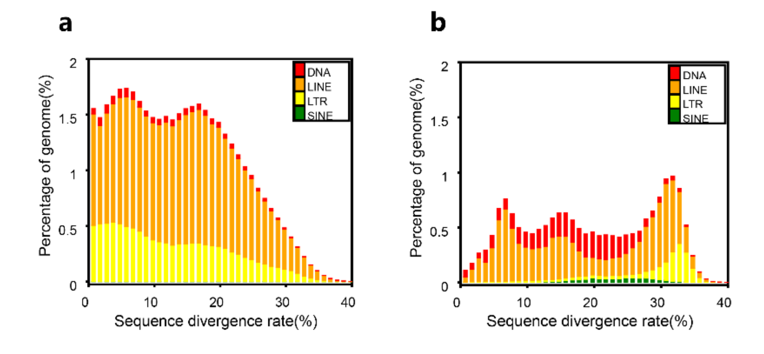


图 3 尖吻蝮基因组转座元件（transposable elements ，Tes）的分布。TEs 包括 DNA 和 RNA 转座子（即 DNAs、LINEs、LTRs 和 SINEs）。**(a)** *de novo* 序列的差异率分布。**(b)** 已知序列的差异率分布。

Figure 3. Distribution of transposable elements（Tes）in the *D. acutus* genome. The TEs include DNA and RNA transposons (i.e., DNAs, LINEs, LTRs and SINEs). **(a)** Divergence rate distribution of the *de novo* sequences. **(b)** Divergence rate distribution of known sequences.

表 3 尖吻蝮基因组重复序列（*de novo*）统计

Table 3 Statistics for the repetitive sequences (*de novo*) from our *D. acutus* genome.

Type	Length (Bp)	% in genome
DNA	27,712,037	1.762318
LINE	464,343,121	29.529418
SINE	984,426	0.062604
LTR	188,498,215	11.987348
Other	0	0
Satellite	1,180,615	0.07508
Simple_repeat	2,250,205	0.143099
Unknown	2,609,514	0.165949
Total	636,067,480	40.45005

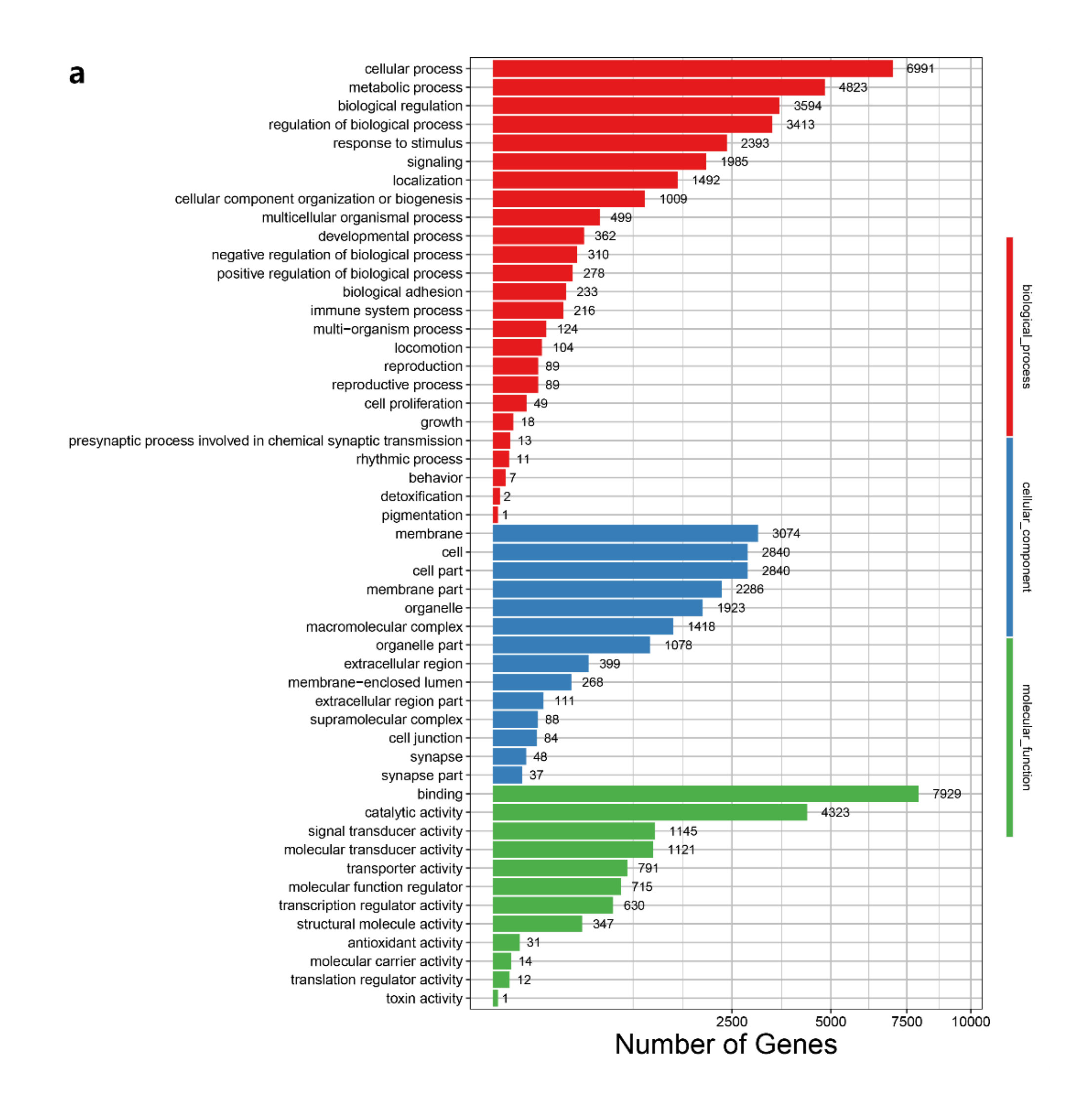
共有 24,402 个功能基因被注释（如表 4 所示）。对功能基因进行基因本体论

（gene ontology ，GO）富集分析结果显示，基因富集于生物过程（biological processes ，BP)、细胞成分（cellular components ，CC）和分子功能（molecular functions ，MF）中。其中细胞过程（cellular process)、膜（membrane）和结合（binding）分别在 BP、CC 和 MF 中含量最高。对功能基因进行 KEGG 通路富集分析结果表明，信号转导相关基因在尖吻蝮中具有至关重要的作用（如图 4 所示)。此外，与代谢相关的富集通路数量最多。

表 4 尖吻蝮基因组功能注释结果

Table 4 Functional annotation result of our *D. acutus* genome.

	Number	Percentage (%)
Total	24,402	100%
Swiss-Prot annotated	19,527	80.02%
KEGG annotated	20,869	85.52%
TrEMBL annotated	22,927	93.96%
InterPro annotated	23,089	94.62%
GO annotated	14,512	59.47%
Overall	23,844	97.71%


a
cellular process 6991
metabolic process 4823
biological regulation 3594
regulation of biological process 3413
response to stimulus 2393
signaling 1985
localization 1492
cellular component organization or biogenesis 1009
multicellular organismal process 499
developmental process 362
negative regulation of biological process 310
positive regulation of biological process 278
biological adhesion 233
immune system process 216
multi−organism process 124
locomotion 104
reproduction 89
reproductive process 89
cell proliferation 49
growth 18
presynaptic process involved in chemical synaptic transmission 13
rhythmic process 11
behavior 7
detoxification 2
pigmentation 1
membrane 3074
cell 2840
cell part 2840
membrane part 2286
organelle 1923
macromolecular complex 1418
organelle part 1078
extracellular region 399
membrane−enclosed lumen 268
extracellular region part 111
supramolecular complex 88
cell junction 84
synapse 48
synapse part 37
binding 7929
catalytic activity 4323
signal transducer activity 1145
molecular transducer activity 1121
transporter activity 791
molecular function regulator 715
transcription regulator activity 630
structural molecule activity 347
antioxidant activity 31
molecular carrier activity 14
translation regulator activity 12
toxin activity 1
biological_process
cellular_component
molecular_function
2500
5000
7500
10000
Number of Genes

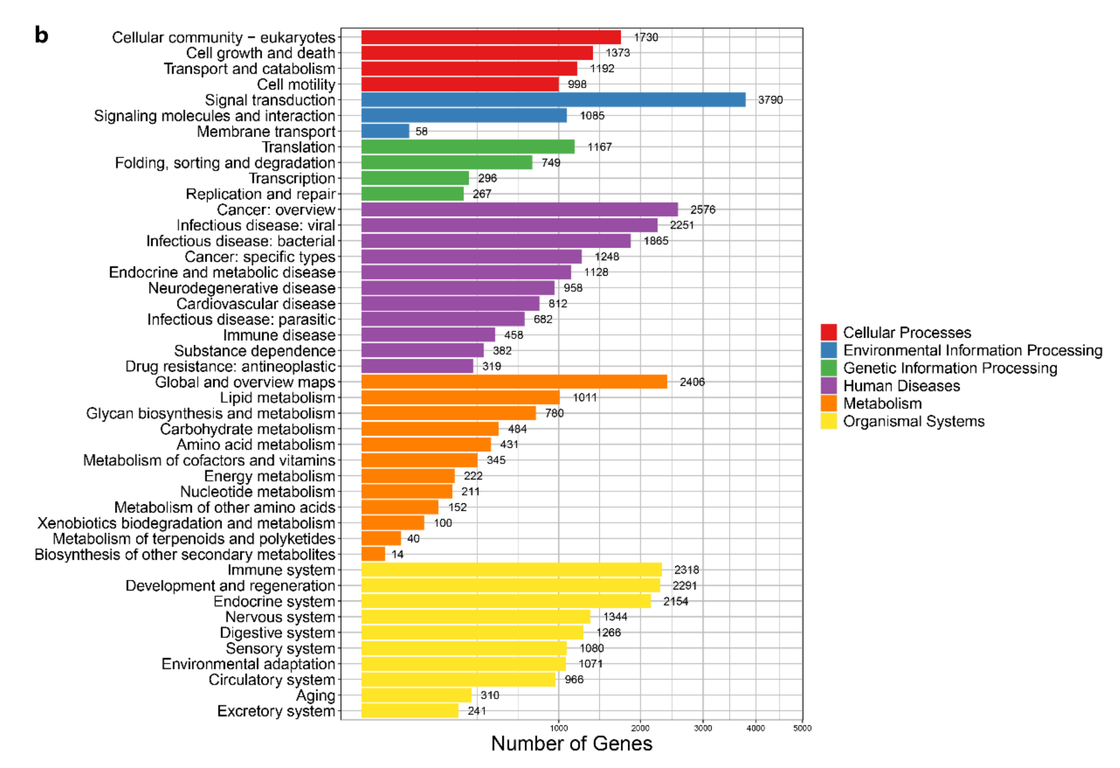

b
Cellular community − eukaryotes 1730
Cell growth and death 1373
Transport and catabolism 1192
Cell motility 998
Signal transduction 3790
Signaling molecules and interaction 1085
Membrane transport 58
Translation 1167
Folding, sorting and degradation 749
Transcription 296
Replication and repair 267
Cancer: overview 2576
Infectious disease: viral 2251
Infectious disease: bacterial 1865
Cancer: specific types 1248
Endocrine and metabolic disease 1128
Neurodegenerative disease 958
Cardiovascular disease 812
Infectious disease: parasitic 682
Immune disease 458
Substance dependence 382
Drug resistance: antineoplastic 319
Global and overview maps 2406
Lipid metabolism 1011
Glycan biosynthesis and metabolism 780
Carbohydrate metabolism 484
Amino acid metabolism 431
Metabolism of cofactors and vitamins 345
Energy metabolism 222
Nucleotide metabolism 211
Metabolism of other amino acids 152
Xenobiotics biodegradation and metabolism 100
Metabolism of terpenoids and polyketides 40
Biosynthesis of other secondary metabolites 14
Immune system 2318
Development and regeneration 2291
Endocrine system 2154
Nervous system 1344
Digestive system 1266
Sensory system 1080
Environmental adaptation 1071
Circulatory system 966
Aging 310
Excretory system 241
1000
2000
3000
4000
5000
Number of Genes
Cellular Processes
Environmental Information Processing
Genetic Information Processing
Human Diseases
Metabolism
Organismal Systems

图 4 尖吻蝮基因组的基因注释。**(a)** GO 富集。**(b)** KEGG 富集。

Figure 4. Gene annotation of our *D. acutus* genome. **(a)** GO enrichment. **(b)** KEGG enrichment.

构建系统发育树结果（如图 5 所示）表明，我们的数据可以用于物种系统发育树的构建且该树与其他人的研究结果一致 [14]。通过将本研究中组装的基因组数据与染色体水平的尖吻蝮基因组数据 [1] 比较可知，我们成功组装并注释一个高度连续的尖吻蝮基因组。

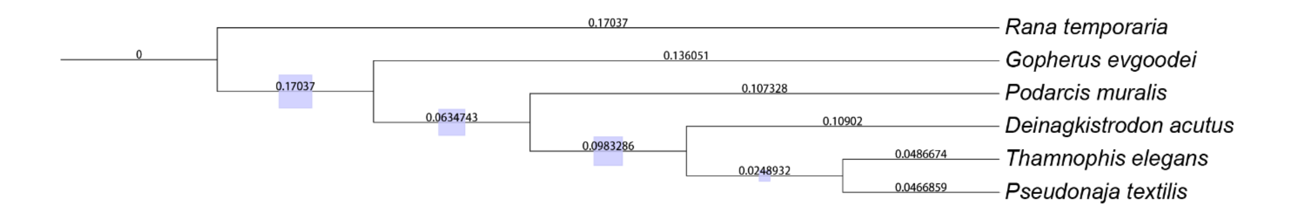


图 5 利用核基因组单拷贝基因重建系统发育树。数字表示分支的长度。有色方块表示 bootstraps/metadata，显示范围为 0.49744 到 1。

Figure 5. Phylogenetic tree reconstructed using single-copy genes from nuclear genomes. The numbers represent the branch lengths. The colored squares represent bootstraps/metadata. The display range is 0.499744 to 1.

重用潜力

我们的数据可为其他人研究尖吻蝮提供参考基因组。此外，它也可与其它蛇类基因组结合使用，用于研究蛇类的系统发育和进化。最后，我们的基因组可为蛇毒和相关毒理学研究提供数据支持。

发表同意

不适用

数据可用性

支持本研究结果的数据已存入国家基因库数据库（CNGBdb）[15] 的 CNGB 序列档案（CNSA）[16]，编号为 CNP0004047。原始数据也可通过 PRJNA955401 在 SRA 中获得。其它数据可在 GigaDB 存储库中获得 [17]。

缩略词表

BP, biological process; CC, cellular component; GO, gene ontology; LINE, long interspersed nuclear element; LTR, long terminal repeat; MF, molecular function; SINE, short interspersed nuclear elements; stLFR, single-tube long fragment read; TE, transposable elements; WGS, Whole Genome Sequencing.

声明

伦理批准

样本采集及相关实验流程经华大基因机构审查委员会批准（BGI-IRB E22017）。

竞争性利益

作者声明没有经济利益冲突。

资助
本项目由广东省高通量基因组测序与合成编辑应用重点实验室资助（grant no. 2017B030301011），本研究也得到国家基因库（CNGB）的支持。

作者贡献
曹军、卢浩荣和刘力榕设计并发起了该项目。蛇的样本由安徽师范大学提供。朱文标和杨淑慧对采集的样本进行处理。王心雨、施敏辉、汪诗晴进行 DNA 提取、文库构建。王心雨进行数据分析并撰写论文。所有作者都已阅读并同意最终手稿。

参考文献

1. Yin W, Wang Z-j, Li Q-y et al. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. Nature communications, 2016; 7(1): 13107. doi:10.1038/ncomms13107.
2. Tan KY, Shamsuddin NN, Tan CH. Sharp-nosed Pit Viper (*Deinagkistrodon acutus*) from Taiwan and China: A comparative study on venom toxicity and neutralization by two specific antivenoms across the Strait. Acta tropica, 2022; 232: 106495. doi:10.1016/j.actatropica.2022.106495.
3. Huang J, Zhao M, Xue C et al. Analysis of the Composition of *Deinagkistrodon acutus* Snake Venom Based on Proteomics, and Its Antithrombotic Activity and Toxicity Studies. Molecules, 2022; 27(7): 2229. doi:10.3390/molecules27072229.
4. Huang F, Zhao S, Tong F et al. Unexpected death in a young man associated with a unilateral swollen leg: Pathological and toxicological findings in a fatal snakebite from *Deinagkistrodon acutus* (Chinese moccasin). Journal of forensic sciences, 2021; 66(2): 786-792. doi:10.1111/1556-4029.14622.
5. Wang D-Q, Pan L-L, Yang D-C et al. Complete mitochondrial genome of the sharp-snouted pitviper *Deinagkistrodon acutus* (Reptilia, Viperidae). Mitochondrial DNA. Part B, Resources, 2019; 4(2): 2900-2901. doi:10.1080/23802359.2019.1660593.
6. Hu X-Q, Wu Q-L, Li X-Y et al. Study on venom protein components of *Deinagkistrodon acutus* living in different geographical units. Oxidation Communications, 2016; 39(A2): 1885-1895.
7. Linfeng W, Lutao X, Pin L et al. Radial artery aneurysm formation and spontaneous rupture after snake bite to the right forearm. Toxicon : official journal of the International Society on Toxinology, 2020; 181: 79-81. doi:10.1016/j.toxicon.2020.04.098.
8. Huang J, Song W, Hua H et al. Antithrombotic and anticoagulant effects of a novel protein isolated from the venom of the *Deinagkistrodon acutus* snake. Biomedicine & Pharmacotherapy, 2021; 138: 111527. doi:10.1016/J.BIOPHA.2021.111527.
9. Huang Z, He D, Liao M. Antibacterial activity of venoms from Guangxi cobra, *Bungarus multicinctus* and *Deinagkistrodon acutus* in vitro. Chinese Journal of Microecology, 2019; 31(10): 1135-1139. doi:10.13381/j.cnki.cjm.201910004.
10. Liu B, Cui L, Deng Z et al. Protocols for the assembly and annotation of snake genomes V.2. 2023; https://dx.doi.org/10.17504/protocols.io.5jyl8j6e9g2w/v2.
11. Liu B, Shi Y, Yuan J et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. aRxiv. 2013; https://doi.org/10.48550/arXiv.1308.2012.

12. Liu B, Cui L, Deng Z et al. The genome assembly and annotation of the many-banded krait, *Bungarus multicinctus*. GigaByte; 2023: gigabyte82. doi:10.46471/gigabyte.82.
13. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology, 2015; 16(1): 157. doi:10.1186/s13059-015-0721-2.
14. Vidal N, Hedges SB. The molecular evolutionary tree of lizards, snakes, and amphisbaenians. Comptes rendus biologies, 2009; 332(2): 129–139. doi:10.1016/j.crvi.2008.07.010.
15. Feng ZC, Li JY, Fan Y et al. CNGBdb: China National GeneBank DataBase. Yi Chuan (Hereditas), 2020; 42(8): 799–809. doi:10.16288/j.yczz.20-080.
16. Guo X, Chen F, Gao F et al. CNSA: a data repository for archiving omics data. Database, 2020; 2020: baaa055. doi:10.1093/database/baaa055.
17. Wang X, Liu L, Zhu W et al. Supporting data for "Genome assembly and annotation of the Sharp-nosed Pit Viper *Deinagkistrodon acutus* based on next-generation sequencing data". GigaScience Database, 2023; http://dx.doi.org/10.5524/102426.